Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Special Track on AI4Tech: AI Enabling Critical Technologies

Eye-See-You: Reverse Pass-Through VR and Head Avatars

Ankan Dash[1] , **Jingyi Gu**[1] , **Guiling Wang**[1] , **Chen Chen**[2]
[1]New Jersey Institute of Technology
[2]University of Central Florida
{ad892, jg95, gwang}@njit.edu, chen.chen@crcv.ucf.edu

Abstract

Virtual Reality (VR) headsets, while integral to the evolving digital ecosystem, present a critical challenge: the occlusion of users' eyes and portions of their faces, which hinders visual communication and may contribute to social isolation. To address this, we introduce **RevAvatar**, an innovative framework that leverages AI methodologies to enable reverse pass-through technology, fundamentally transforming VR headset design and interaction paradigms. RevAvatar integrates state-of-the-art generative models and multimodal AI techniques to reconstruct high-fidelity 2D facial images and generate accurate 3D head avatars from partially observed eye and lower-face regions. This framework represents a significant advancement in AI4Tech by enabling seamless interaction between virtual and physical environments, fostering immersive experiences such as VR meetings and social engagements. Additionally, we present **VR-Face**, a novel dataset comprising 200,000 samples designed to emulate diverse VR-specific conditions, including occlusions, lighting variations, and distortions. By addressing fundamental limitations in current VR systems, RevAvatar exemplifies the transformative synergy between AI and next-generation technologies, offering a robust platform for enhancing human connection and interaction in virtual environments.

1 Introduction

Augmented Reality (AR) and Virtual Reality (VR) have become critical technological advancements, transforming industries such as gaming, remote collaboration, education, and healthcare [Al-Ansi *et al.*, 2023; Rambach *et al.*, 2020; Kanschik *et al.*, 2023]. As immersive technologies, they enable new forms of interaction and engagement, reshaping human-computer interfaces and digital experiences. While VR headsets have become mainstream consumer technology, they inherently isolate users from their surroundings, limiting their integration into shared environments and public spaces [Hobbs, 2017; Gugenheimer *et al.*, 2019]. Eye contact is a cornerstone of human connection and emotional communication, yet current VR headsets obscure users' eyes and facial expressions, severing visual interaction with the real world. This lack of transparency not only diminishes social presence but also leaves bystanders unaware of the user's engagement with VR content or their attentiveness.

Addressing this fundamental limitation requires transformative AI-driven solutions to bridge the gap between virtual and physical environments. One such approach is *reverse pass-through* technology, which reconstructs and displays a user's eyes and facial expressions on the outward-facing surface of the headset. This technique enables real-time interaction, allowing bystanders to perceive eye movements and emotional expressions, effectively bridging the gap between virtual and physical environments.

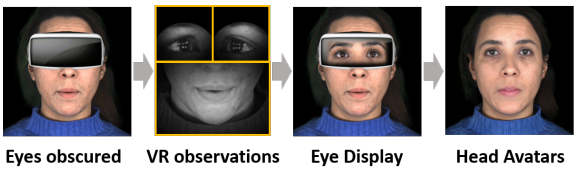

Figure 1: Our proposed RevAvatar framework for reverse pass-through, enabling the display of eyes and full-head avatars.

Efforts to mitigate VR isolation made effort to maintain social presence, while they suffer from inauthentic eye movements [Chan and Minamizawa, 2017; Bozgeyikli and Gomes, 2022], hardware requirements, limited facial reconstruction [Matsuda *et al.*, 2021], and underwhelming performance [Apple, 2024], as shown in Table 1. Besides, photo-realistic avatar generation has significantly improved realism [Feng *et al.*, 2021; Danecek *et al.*, 2022; Zheng *et al.*, 2023; Grassal *et al.*, 2022; Li *et al.*, 2023], but limited by the need for multi-view images or specialized VR headsets. Person-specific avatar generation struggles with high customization needed, hindering widespread adoption [Lombardi *et al.*, 2018; Wei *et al.*, 2019; Schwartz *et al.*, 2020].

To address the limitations of current VR systems, we propose RevAvatar, a framework to generalize across VR headsets with minimal device-specific fine-tuning. Unlike existing solutions, **RevAvatar** leverages advanced AI techniques to overcome the isolation caused by VR headsets by reconstructing and displaying the user's eyes and facial expressions in

	Methods	Pros	Cons
Reverse Pass	FrontFace, Google Eyes [Matsuda *et al.*, 2021]	Displays animated eyes Improves social presence	Lacks authenticity; fails to convey emotions Requires custom hardware; lacks full facial reconstruction
	EyeSight-Apple Vision Pro	Displays eyes externally	Limited realism; unclear effectiveness
	Ours	Full face reconstruction	-
Avatar	Photo-realistic Avatar	Increases realism	Requires multi-view images and specialized VR headsets
	Person-specific Avatar	Enables personalization	High customization limits scalability
	Ours	One-shot avatar	-

Table 1: Comparison of Existing VR Social Presence Methods

real-time, enabling seamless interaction between virtual and physical environments. Additionally, RevAvatar facilitates the creation of full-head 3D avatars, enhancing immersive experiences for applications like virtual meetings. The main pipeline combines real-time 2D face restoration for "reverse pass-through" and a one-shot 3D avatar generation model, achieving 0.008-second inference on mobile SoCs like Apple M2. It is compatible with consumer mixed-reality devices such as the Apple Vision Pro and upcoming Samsung and Google VR headsets. Crucially, it eliminates the need for 3D scans, requiring only a selfie-like Digital Persona (DP) image, improving accessibility and convenience.

Developing generalized solutions for diverse VR headsets is challenging due to variability in camera specifications and placements across brands like Apple, Samsung, Meta, Vive, and Varjo. To address this, we introduce **VR-Face**, a novel dataset comprising 200,000 samples that simulate diverse VR conditions, including occlusions, lighting variations, and distortions. VR-Face not only supports RevAvatar's development but also provides a foundational resource for advancing research in AI-driven VR technologies.

Our contributions are: ① **RevAvatar Framework**: We introduce **RevAvatar**, an AI-driven framework for real-time reverse pass-through and 3D avatar generation in VR. This solution enhances VR immersion by eliminating the need for user-specific models or custom hardware. ② **Efficient AI for Mobile SoCs**: Our 2D face reconstruction model operates efficiently on mobile SoCs, such as the Apple M2 in Apple Vision Pro, achieving an inference time of just 0.008 seconds. This demonstrates its scalability across diverse VR devices. ③ **VR-Face Dataset**: We present **VR-Face**, a public dataset with 200,000 samples simulating challenging VR scenarios. It supports the development of AI technologies adaptable to various headset specifications and advances research in VR. ④ **AI-Enabled VR Advancements**: Through RevAvatar and VR-Face, we drive significant AI innovations that address key VR challenges like user isolation and hardware diversity, setting a new standard for AI-driven progress in VR.

2 Related Work

Eye tracking based animation FrontFace [Chan and Minamizawa, 2017] and Googly Eyes [Bozgeyikli and Gomes, 2022] aimed to represent user attention and gaze using animated eye movements via eye tracking and Head-Mounted Displays (HMDs). However, these approaches focus on displaying *animated* eye states, such as whether the eyes are open or closed, and the gaze direction, without conveying genuine emotions or expressions.

Eye and Face Reconstruction: Reverse Pass-Through A "reverse pass-through" prototype headset [Matsuda *et al.*, 2021] reconstructs and displays users' eyes on an external screen but requires costly and custom hardware inaccessible to most users. Apple's Vision Pro with "EyeSight" projects eyes onto an external display, but its functionality remains unclear [Apple, 2024], and early reviews suggest underwhelming performance [Chokkattu, 2024; Patel, 2024]. Other face-restoration methods from partial VR headset data often require multiple views or customized headsets, making them impractical for widespread use. They also rely on user-specific avatars, requiring costly individualized training and limiting real-world use [Lombardi *et al.*, 2018; Wei *et al.*, 2019; Schwartz *et al.*, 2020].

Face image composition and one-shot Avatar generation Several methods were proposed for face composition and synthesis. PixelStyle2Pixel (PSP) [Richardson *et al.*, 2021] performs image-to-image translation using StyleGAN's latent space, while failing to preserve the identity of unseen individuals. StyleMapGAN [Kim *et al.*, 2021] faces similar identity preservation challenges during inference. One-shot photo-realistic avatar generation made significant strides, like ROME [Khakhulin *et al.*, 2022] generating mesh-based avatars from single images, and CVTHead [Ma *et al.*, 2024] using transformers and point-based neural rendering. Portrait4D [Deng *et al.*, 2024] employs a part-wise 4D generative model for synthesizing multi-view images and leverages transformers to create highly detailed, animatable avatars. They are closest to our task of one-shot avatar generation.

VR simulation dataset Despite growing interest in VR simulations, publicly available datasets remain scarce. Eye images captured by IR cameras in VR headsets suffer from occlusions and limited fields of view. MEAD data [Chen *et al.*, 2024] was used for VR simulations but lacked real-world scenario complexity, particularly in occlusions like eyebrow obstruction. Other works [Lombardi *et al.*, 2018; Wei *et al.*, 2019] used custom VR headsets with IR cameras for optimal eye capture, but these setups are not generalizable to commercial headsets, and datasets are publicly inaccessible.

3 VR-Face Dataset

We develop the *VR-Face* dataset (Figure 2), containing 200,000 samples, each comprising a full-face image, left and

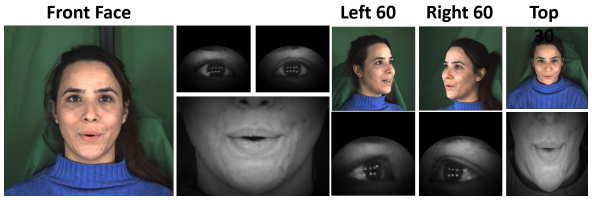

Figure 2: Sample processed images from the *VR-Face* dataset simulating VR environments.

Aspect	Description
Total Samples	200,000
Image Types	Full-face, left eye, right eye, lower face
Angles	Front, left-60°, right-60°, top-30°
Expressions	Anger, contempt, disgust, fear, happy, neutral, sad, surprise
Preprocessing	Distortion, masking, vignetting, blur, grayscale
Effects	Occlusions, lighting shifts, noise, eyebrow reduction
Diversity	Skin tones, genders, ethnicities

Table 2: Overview of the VR-Face Dataset

right eye images from various angles, and a lower-face image. The dataset captures a wide range of facial expressions and perspectives, with pre-processing to simulate visual effects observed in VR headset imagery (details in Table 2). VR-Face is designed to be inclusive, representing diverse skin tones, sex, race, and ethnicity. While it serves as a benchmark for reverse pass-through and analysis, it is not intended to replace real-world VR headset data but can be adapted to specific devices with suitable datasets.

4 Methodology

RevAvatar comprises an initial setup stage and the main pipeline, which consists of image alignment, 2D face restoration, and 3D avatar generation for reverse pass-through and immersive VR experience, as shown in Figure 3. and 4.

During the initial VR headset setup, the user is prompted to capture a selfie using the headset's external camera or upload an image via their online account. Manufacturers provide detailed instructions to ensure the image captures the entire face without occlusions and under good lighting. The image is then processed using a facial landmark model [Bulat and Tzimiropoulos, 2017] for cropping and alignment in subsequent stages. This selfie, referred to as the Digital Persona (DP) image, serves as the reference for 2D and 3D reconstruction.

The main pipeline aligns or frontalizes the processed eye and face-tracking images. A lightweight GAN-based model [Goodfellow *et al.*, 2014; Dash *et al.*, 2024] is then used for full face restoration. For Avatar generation, we combine 3DMM [Egger *et al.*, 2020] models with neural networks for tri-plane representation and volume rendering to achieve accurate 3D reconstruction.

4.1 Eye and Face Alignment

We use VR-Face dataset, consisting of non-aligned images of left and right eyes, with the goal of aligning or frontalizing these images (Figure 3-A). To achieve this, we adopt a CycleGAN-based framework [Zhu *et al.*, 2017], which learns bidirectional mappings between tilted eye images (domain A) and frontal eye images (domain B) through Cycle-Consistency. We use two distinct alignment models, CG_{LE} and CG_{RE} for each eye, and use a separate model CG_{Face} to frontalize the lower face image. To preserve gaze, we use eye landmark detection [Bulat and Tzimiropoulos, 2017] and compute Gaze Estimation Error as the angular difference between aligned and ground truth images, achieving errors below one degree. Aligned eye and lower face images are then pasted onto the DP image using facial landmarks. Dataset details are in Section 5.1, with examples in Figure 3-B.

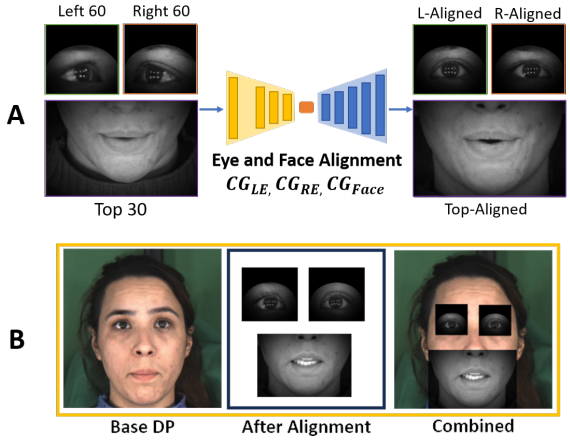

Figure 3: A - Left, Right Eye and Face alignment models based on CycleGAN. B - The combined image is used as input for full face restoration.

4.2 2D Full Face Restoration

To restore a full face image from frontalized eye and face tracking images, we develop a lightweight GAN-based restoration model with a Generator G and Discriminator D. It takes aligned grayscale left and right eye images, along with a lower face image pasted onto the color DP image, and generates a restored output where the grayscale images blend seamlessly with the rest of the face, capturing facial expressions. Figure 4. shows our 2D Face Restoration Framework. To handle high occlusion of input eye images, the model uses a reference image for reconstructing facial features, providing context for occluded areas like eyebrows. During training, random reference images from various users are selected to improve robustness. During deployment, the DP image serves as a reference for reconstructing occluded areas.

The Generator G comprises an Input Encoder E_I and a Reference Encoder E_R, which share a similar structure. E_I takes the DP image with partial VR observations pasted on top of the base DP image x as the input, and the reference DP image z is given as input to E_R. To enhance the model's ability to capture both global and fine-grained details, the encoders are extended to operate at multiple scales, where each scale extracts features at different levels of resolution. The Generator is built on multiple ResNet blocks, with each block processing features at different resolutions. We use

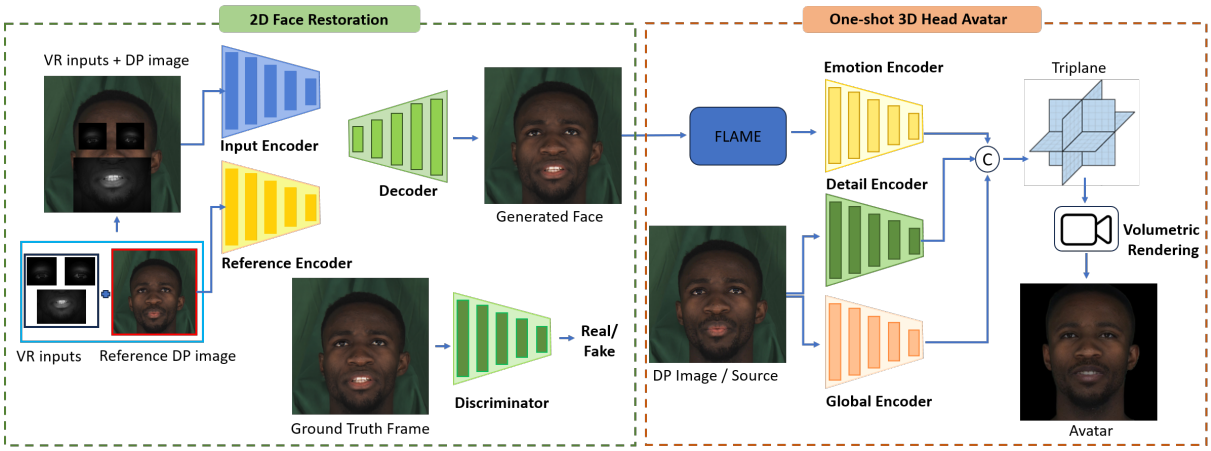

Figure 4: Overview of the **2D Face Restoration** and **One-Shot Avatar** generation model. **2D Face Restoration:** Partial VR observations and the reference DP are inputs to the Input Encoder, while the reference image is processed by the Reference Encoder. The restored face output drives the one-shot avatars. **One-Shot Avatar:** The DP image serves as the source, and the restored image from the 2D face restoration model drives the avatar generation. A tri-plane is generated from concatenated encoder outputs, followed by volumetric rendering and super-resolution to produce the final output.

cross-attention module to align and integrate features from the reference image with those of the input image, enhancing context-aware and guided image generation. The cross-attention module operates at multiple scales, where each scale learns to focus on different levels of detail, from coarse structures to fine textures. This multi-scale feature fusion ensures the generation of high-fidelity images by combining both global context and local detail. The architecture begins with initial convolutional layers for downsampling through E_I and E_R, extracting multi-scale features. The multi-scale features are then processed through cross-attention and residual block processing. Finally, the decoder reconstructs the synthesized image $G(x, z)$, utilizing the multi-scale features to improve the quality of the generated image at all levels. We use a Multiscale Discriminator, which employs multiple instances of a PatchGAN [Isola *et al.*, 2017] discriminator, each responsible for evaluating the image at a specific scale.

The Generator G is trained to minimize the following loss functions. (1) The adversarial loss $\mathcal{L}_{adv}$ ensures that the generated image $G(x, z)$ is indistinguishable from real images y by the Discriminator D. (2) The L1 loss $\mathcal{L}_{L1}$ ensures the generated image $G(x, z)$ is close to the ground truth image y. (3) The LPIPS (Learned Perceptual Image Patch Similarity) [Zhang *et al.*, 2018] loss $\mathcal{L}_{LPIPS}$ assesses perceptual similarity between the generated image $G(x, z)$ and the target image y. The total loss is defined as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{G}} + \mathcal{L}_{\text{D}}$.

$$\mathcal{L}_G = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{L1}\mathcal{L}_{L1} + \lambda_{LPIPS}\mathcal{L}_{LPIPS} \quad (1)$$

$$\mathcal{L}_{adv} = \mathbb{E}_y\left[\log D(y)\right] + \mathbb{E}_x\left[\log(1 - D(G(x, z)))\right] \quad (2)$$

where the weights λ_{adv}, λ_{L1}, and λ_{LPIPS} balance the contributions of each loss term.

The Discriminator D is trained to distinguish between real images and generated images using the adversarial loss:

$$\mathcal{L}_D = \mathbb{E}_y\left[\log D(y)\right] + \mathbb{E}_x\left[\log(1 - D(G(x, z)))\right] \quad (3)$$

where $D(y)$ is the probability that the image y is real, and $D(G(x, z))$ is the probability that the generated image $G(x, z)$ is real.

4.3 One-shot 3D Head Avatar Model

We extend our reverse pass-through system to generate full-head avatars for immersive VR, building on recent one-shot facial avatar generation advancements. This approach overcomes the limitations of requiring specialized models for each subject or multi-view inputs, addressing challenges that hinder practical real-world applications.

Our Framework uses the user's DP or selfie image as the source image I_s and the reconstructed image from 2D full-face restoration serves as the driving or target image I_t. The source image is used to extract the identity, and the target image is responsible for providing the pose and expression information. Our framework comprises three main branches which include the Global Branch E_G, Detail Branch E_D and the Expression Branch E_E. The output is then up-scaled and refined using a super-resolution module.

The E_G branch uses a hybrid transformer model with a series of convolutional and transformer blocks along with SegFormer [Xie *et al.*, 2021] to generate a tri-plane representation. We use SegFormer as it allows for effective mapping from 2D space to 3D space. This is achieved by predicting a tri-plane T_g that represents the neutral expression of the human face in a canonical 3D space. To ensure that the generated tri-plane T_g aligns with the identity of I_s and maintains a neutral expression, we incorporate a 3D Morphable Model (3DMM) to render a face with the same identity and camera pose as the source image, but with a neutral expression. The Detail Branch E_D builds on the geometry provided by the Global Branch by capturing and reconstructing intricate facial details from the source image I_s. The features of the Detail Branch are transferred to the global triplane, creating an appearance triplane T_d that improves the initial reconstruction with fine-grained details, such as texture and surface features. The Expression Branch focuses on modeling and transferring the expression from the target image I_t onto the reconstructed 3D avatar. This branch utilizes a 3DMM to predict the shape

and expression coefficients for both the source image I_s and target image I_t. The expression coefficients of I_t are used to render a frontal-view expression image I_e, which is then encoded into an expression tri-plane T_e. This expression tri-plane is added to the canonical tri-plane T_g along with the appearance tri-plane T_d to generate the final 3D reconstruction. The integration of these three branches allows the model to combine the identity from the source image with the expression and head pose from the target image, effectively transferring the desired expression onto the source image while maintaining high fidelity in both appearance and geometry. Given the high computational demands of volumetric rendering, we first render low-resolution images and then use a super-resolution [Wang *et al.*, 2021] module to produce the final high-quality output.

We use a two-stage training schedule for multi-view consistency and efficiency. In the first stage, the model trains at a lower resolution without an upscaling module, optimizing L_1 and L_{LPIPS} losses between the Global Branch feature rendering and the 3DMM rendering of the source image, as well as the combined tri-plane features and target image.

$$\begin{aligned}\mathcal{L}_{\text{G}} &= L_1(R(T_{\text{g}}), R_{\text{3DMM}}(I_{\text{s}})) \\ &\quad + L_{\text{LPIPS}}(R(T_{\text{g}}), R_{\text{3DMM}}(I_{\text{s}})) \qquad (4)\\ \mathcal{L}_{\text{Combined}} &= L_1(R(T_{\text{combined}}), I_{\text{t}}) \\ &\quad + L_{\text{LPIPS}}(R(T_{\text{combined}}), I_{\text{t}}) \qquad (5)\\ \mathcal{L}_{\text{Stage1}} &= \lambda_{\text{G}}\mathcal{L}_{\text{G}} + \lambda_{\text{Combined}}\mathcal{L}_{\text{Combined}} \qquad (6)\end{aligned}$$

In the second stage of training, we only fine-tune the upscaling module using L_1, L_{LPIPS}, and GAN loss objective. Additionally, we use an eye region loss which calculates the L_1 between only the eye region rendering the output image and the target ground truth image to ensure accurate gaze.

$$\begin{aligned}\mathcal{L}_{\text{Stage2}} &= \lambda_{\text{L1}}L_1(I_{\text{o}}, I_{\text{t}}) + \lambda_{\text{LPIPS}}L_{\text{LPIPS}}(I_{\text{o}}, I_{\text{t}}) \\ &\quad + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}} + \lambda_{\text{Eye}}L_1(I_{\text{eye-output}}, I_{\text{eye-target}}) \\ &\quad + L_{\text{LPIPS}}(I_{\text{eye-output}}, I_{\text{eye-target}}) \qquad (7)\\ \mathcal{L}_{\text{Total}} &= \mathcal{L}_{\text{Stage1}} + \mathcal{L}_{\text{Stage2}} \qquad (8)\end{aligned}$$

4.4 Reverse Pass-through and Avatar Outputs

The output from our Full Face Restoration model can be cropped to display the eye region and realize reverse pass-through on mainstream VR headsets(Figure 5). This allows users to maintain eye contact and convey expressions. Additionally, the outputs from our head avatar model can be leveraged for immersive applications such as VR meetings and the metaverse, providing visually accurate and expressive 3D avatars that enhance virtual interactions and communication.

5 Experiments

5.1 Datasets

We utilize *VR-Face* as the main dataset to train and test our framework, and three additional datasets for training to enhance the generalization. (1) The Eye and Face Alignment model is exclusively trained and tested on *VR-Face* dataset. (2) For 2D face restoration model, we integrate CelebHQ [Karras *et al.*, 2017] and FFHQ [Karras *et al.*, 2019] datasets, which contain images only, in conjunction with *VR-Face* dataset for training, allowing for better generalization across different skin tones and facial attributes. (3) For 3D avatar generation, besides *VR-Face*, we further leverage FFHQ, CelebV-HQ [Zhu *et al.*, 2022], VFHQ [Xie *et al.*, 2022] datasets to provide a rich set of facial attributes and emotional variations.

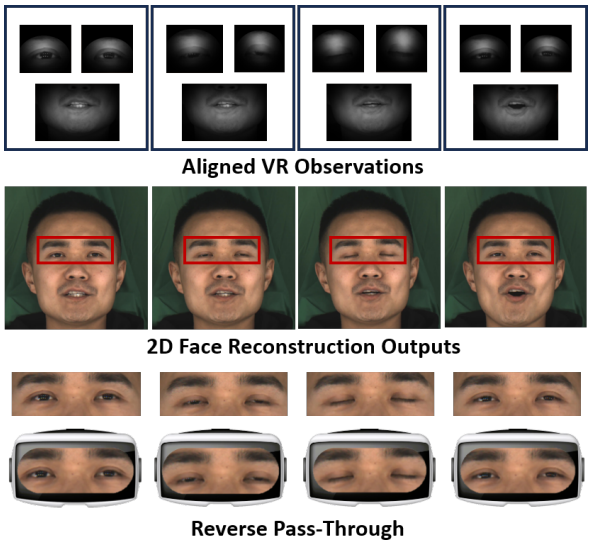

Figure 5: Sample output from 2D Face Restoration model which can be used for Reverse pass-through.

5.2 Implementation

All models are trained on 512×512 images. The Alignment and 2D face restoration models use a single A100 GPU, while the 3D head avatar model is trained on 4 A100 GPUs. We train three alignment models: two for the eyes and one for the lower face. Batch sizes are set to 32 for Alignment, 16 for 2D face restoration, and 4 for the 3D avatar model. All models use the Adam optimizer with a 0.0001 learning rate. During inference, the alignment models run in 0.004s, the face restoration model in 0.006s per image, and the avatar model achieves 22 FPS. All inferences are performed on a single A100 GPU including mobile GPUs such as Apple M2 GTX 1050, and MX350.

5.3 Baseline

Given the lack of well-established baseline models for reverse pass-through VR, a direct holistic comparison of our entire framework is not feasible. Instead, we evaluate our 2D full-face restoration model by comparing it with state-of-the-art GAN-based approaches in image composition and reconstruction: CycleGAN, PSP [Richardson *et al.*, 2021], and SMG [Kim *et al.*, 2021]. Additionally, we include DifFace [Yue and Loy, 2024] as a diffusion-based baseline. Although we considered diffusion models such as OSDFace [Wang *et al.*, 2024], DifFace [Yue and Loy, 2024], OSEDiff [Wu *et al.*, 2024], and DiffBIR [Lin *et al.*, 2024], **their inference times of 0.1, 6.1, 0.12, and 8.01 seconds, respectively, on an A100 GPU make them unsuitable for real-time applications**. In contrast, CycleGAN, PSP, SMG, and our model

Model	Full Face			Eye Region of Interest			Inference Time (s)
	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	
CycleGAN	0.8414	23.0429	0.0618	0.6711	20.2291	0.1122	0.016
PSP	0.6271	19.6809	0.1714	0.5737	18.7591	0.1694	0.041
SMG	0.6521	21.0219	0.2349	0.6211	19.2129	0.1521	0.039
DifFace (2024)	**0.9541**	29.8129	0.1306	0.8122	26.0021	0.1023	6.125
Ours	0.9445	**31.3951**	**0.0243**	**0.8572**	**28.2897**	**0.0510**	**0.006**

Table 3: Quantitative Comparison for Full Face and Eye Region Reconstruction, including Inference Time (s).

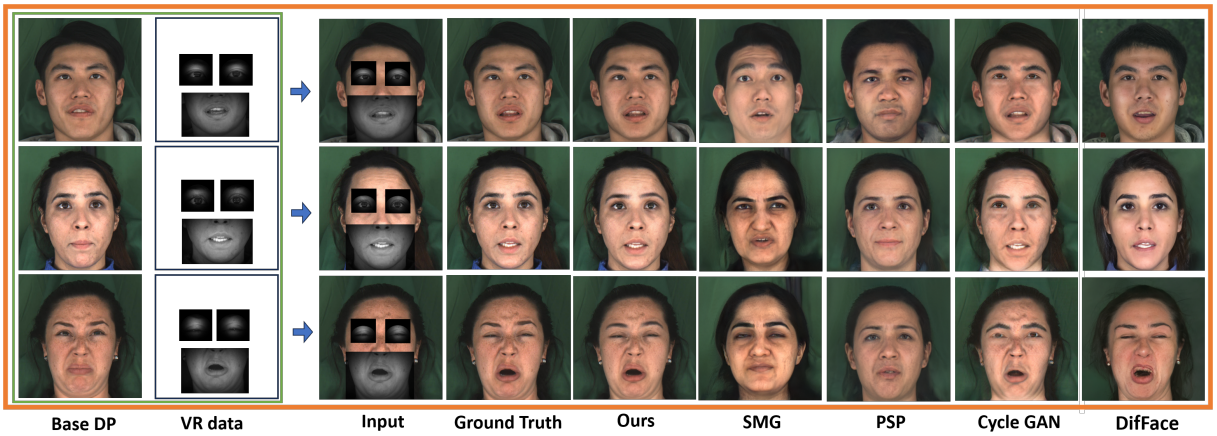

Figure 6: Qualitative comparison of full-face restoration results on unseen test data.

Model	Inference Time (s)			
	A100	Apple M2	MX350	GTX 1050
CycleGAN	0.016	0.020	0.350	0.278
PSP	0.041	0.050	0.852	0.544
SMG	0.039	0.045	0.791	0.365
Ours	**0.006**	**0.012**	**0.125**	**0.0517**

Table 4: Inference Time for A100, Apple M2, MX350, and GTX 1050 for Full Face Reconstruction for Reverse Pass-Through.

achieve real-time performance with **inference times of 0.016, 0.041, 0.038, and 0.006 seconds, respectively**. For 3D avatar reconstruction, we benchmark our model against leading one-shot approaches: ROME [Khakhulin *et al.*, 2022], CVTHead [Ma *et al.*, 2024], and Portrait-4D [Deng *et al.*, 2024].

5.4 Results

For our 2D face reconstruction model and avatar, we generate images at 512x512 resolution and evaluate them against ground truth images using three metrics: SSIM [Wang *et al.*, 2004], PSNR, and LPIPS. These metrics assess visual accuracy and perceptual quality by considering structural similarity, pixel-level differences, and perceptual relevance.

2D Face Restoration Figure 6. and Table 3. presents the qualitative and quantitative comparison between our face restoration model and baselines. Despite its lightweight design, our model performs better than other baselines. While CycleGAN performs comparably to other GAN-based models, it struggles to effectively colorize and blend eye and lower face, leading to severe artifacts, as shown in Figure 6. PSP and SMG, which rely on StyleGAN-based generators, map inputs to a latent space, resulting in a loss of identity and inaccurate face restorations. During testing, SMG tends to output similar images from its training set but with altered expressions, while PSP produces outputs that often diverge significantly from the ground truth, as highlighted in Figure 6. This exposes a critical limitation of StyleGAN-based models: poor generalization on unseen data. DifFace, as a diffusion-based model, achieves high SSIM and PSNR for full-face reconstruction, outperforming other baselines in preserving global facial structure. However, it struggles to retain individual identity, leading to subtle yet noticeable shifts in facial features. Additionally, the iterative nature of the diffusion process results in significantly higher inference time, making DifFace less suitable for real-time applications. In contrast, our model excels at handling unseen face images, demonstrating superior generalization capabilities. It achieves the highest PSNR and the lowest LPIPS, indicating better perceptual quality and sharpness. Moreover, its inference time is orders of magnitude faster than DifFace, making it highly efficient for real-time applications. Figure 5 depicts how the output of our face restoration model enables real-time reverse pass-through capabilities in VR applications.

3D Head Avatar For 3D avatar generation, we compare rendered images with ground truth images. Additional videos showing different views for the reconstructed 3D avatars, as well as the appendix, are available at[1]. Our model shows significant improvements in key metrics, achieving the highest SSIM and PSNR, along with the lowest LPIPS, as shown in

[1]https://github.com/ankan2709/eye-see-you-vr

Model	SSIM↑	PSNR↑	LPIPS↓
ROME (ECCV'22)	0.7522	22.7538	0.1089
CVTHead (WACV'24)	0.7616	21.5395	0.1368
Portrait4D-v2 (ECCV'24)	0.7922	24.3271	0.0638
Ours	**0.8025**	**25.1284**	**0.0629**

Table 5: Comparison of one-shot avatar models.

Table 5. It indicates that our model excels in maintaining both structural integrity and perceptual quality. The high SSIM reflects our model's ability to accurately capture fine details and facial features, while the superior PSNR highlights its robustness in minimizing reconstruction noise and artifacts. Moreover, the lower LPIPS suggests that our method produces image reconstructions that are perceptually closer to the ground truth, ensuring high-fidelity 3D avatars with realistic texture. Although Portrait4D-v2 performs competitively and ranks slightly behind our model, the noticeable jittering in the output faces during eye blinks affects the overall realism. ROME and CVTHead exhibit more challenges in preserving facial identity, reflected in higher LPIPS and lower SSIM. CVTHead, in particular, struggles to maintain identity consistency across different poses, as shown in Figure 7. In contrast, our approach preserves identity more effectively, yielding visually accurate avatars that are faithful to the subject's original appearance. This underscores our model's ability to generate high-quality, realistic 3D avatars with improved generalization across diverse inputs.

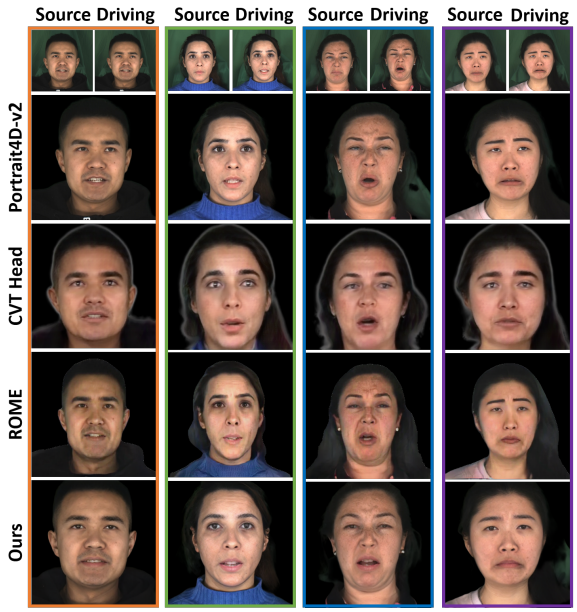

Figure 7: Qualitative comparison of one-shot full head avatar generation.

Realtime Face Reconstruction To assess the real-time performance of our 2D face reconstruction model for reverse pass-through on VR headsets like the Apple Vision Pro (with Apple M2 chip) and Meta Quest 3, we tested the model and compared inference times on the A100 GPU (used for training) with the Apple M2 SoC (in the Vision Pro), tested on a MacBook Air (8-core CPU, 4 performance cores, 4 efficiency cores, 10-core GPU, 16-core Neural Engine, and 16GB unified memory), as well as on the NVIDIA MX350 and GTX1050. Table 4 shows that the Apple M2, with MPS acceleration, delivers inference times comparable to the A100, demonstrating its ability to efficiently run complex models in real-time. The NVIDIA MX350, though slower due to its older architecture, and the GTX1050, showed promising performance for model inference.

Model	SSIM↑	PSNR↑	LPIPS↓
2D Face-Recon (original)	**0.9445**	**31.395**	**0.0243**
AE alignment model	0.8725	27.2211	0.1025
w/o cross attention	0.9124	27.4019	0.0921
w/o LPIPS loss	0.9102	29.0121	0.1001
w/o Reference Image	0.8921	28.209	0.1129

Table 6: Ablation study for framework components.

Ablation Study Table 6 presents the quantitative results of ablation experiments. **Eye alignment ablation:** We compare the performance of Cycle-GAN versus Auto Encoder (AE) [Bank *et al.*, 2021] for eye and face alignment, as described in section 4.1. The results show that Cycle-GAN outperforms AE in alignment tasks, leading to better reconstruction quality. **Cross attention ablation:** The model's performance significantly degrades when the reference and input features are concatenated, rather than using cross attention. This highlights the importance of cross attention in capturing fine-grained details for accurate reconstruction. **Reference image ablation:** Omitting the reference image results in lower performance, as it limits the model's ability to accurately reconstruct occluded areas, which are critical for realistic face restoration. The absence of this contextual information hinders the model's ability to recover facial features that are obstructed in the input. **LPIPS ablation:** Excluding LPIPS loss degrades the perceptual quality of the generated images, as evidenced by increased LPIPS score. Including it helps the model generate more visually accurate and perceptually consistent reconstructions by optimizing for human perception rather than pixel-wise similarity alone.

6 Conclusion

We introduce RevAvatar, an AI-driven solution to mitigate social isolation induced by VR headsets by restoring full-face images from tracking cameras using a user's DP image, enabling real-time eye movement display on an outward-facing VR screen. Additionally, RevAvatar generates realistic one-shot full-head avatars for VR meetings and interactions. As AR/VR continues to revolutionize digital interaction, we support this advancement with VR-Face, a dataset designed to simulate real-world VR scenarios and drive research in this field. Through RevAvatar and VR-Face, we aim to set new benchmarks for AI-driven VR experiences, enhancing social presence and immersion.

References

[Al-Ansi *et al.*, 2023] Abdullah M. Al-Ansi, Mohammed Jaboob, Askar Garad, and Ahmed Al-Ansi. Analyzing augmented reality (ar) and virtual reality (vr) recent development in education. *Social Sciences & Humanities Open*, 8(1):100532, 2023.

[Apple, 2024] Apple. What does eyesight show? https://support.apple.com/guide/apple-vision-pro/about-eyesight-tan5162eaaec/visionos, 2024. Accessed: 2024-12-11.

[Bank *et al.*, 2021] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders, 2021.

[Bozgeyikli and Gomes, 2022] Evren Bozgeyikli and Victor Gomes. Googly eyes: Displaying user's eyes on a head-mounted display for improved nonverbal communication. In *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '22, page 253–260, New York, NY, USA, 2022. Association for Computing Machinery.

[Bulat and Tzimiropoulos, 2017] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[Chan and Minamizawa, 2017] Liwei Chan and Kouta Minamizawa. Frontface: facilitating communication between hmd users and outsiders using front-facing-screen hmds. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '17, New York, NY, USA, 2017. Association for Computing Machinery.

[Chen *et al.*, 2024] Z. Chen, Z. Zhang, J. Yuan, Y. Xu, and L. Liu. Show your face: Restoring complete facial images from partial observations for vr meeting. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8673–8682, Los Alamitos, CA, USA, jan 2024. IEEE Computer Society.

[Chokkattu, 2024] Julian Chokkattu. Review: Apple vision pro, March 28 2024.

[Danecek *et al.*, 2022] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022.

[Dash *et al.*, 2024] Ankan Dash, Junyi Ye, and Guiling Wang. A review of generative adversarial networks (gans) and its applications in a wide variety of disciplines: From medical to remote sensing. *IEEE Access*, 12:18330–18357, 2024.

[Deng *et al.*, 2024] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[Egger *et al.*, 2020] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models – past, present and future, 2020.

[Feng *et al.*, 2021] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG), Proc. SIGGRAPH*, 40(4):88:1–88:13, August 2021.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[Grassal *et al.*, 2022] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022.

[Gugenheimer *et al.*, 2019] Jan Gugenheimer, Christian Mai, Mark McGill, Julie Williamson, Frank Steinicke, and Ken Perlin. Challenges using head-mounted displays in shared and social spaces. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, page 1–8, New York, NY, USA, 2019. Association for Computing Machinery.

[Hobbs, 2017] Thomas Hobbs. Google admits vr is still far too much of an 'isolating' experience, November 07 2017.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017.

[Kanschik *et al.*, 2023] Dominika Kanschik, Raphael Romano Bruno, Georg Wolff, Malte Kelm, and Christian Jung. Virtual and augmented reality in intensive care medicine: a systematic review. *Annals of Intensive Care*, 13(1):81, Sep 2023.

[Karras *et al.*, 2017] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.

[Karras *et al.*, 2019] Tero Karras, S. Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019.

[Khakhulin *et al.*, 2022] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *European Conference of Computer vision (ECCV)*, 2022.

[Kim *et al.*, 2021] Hyunsu Kim, Yunjey Choi, Junho Kim, Sungjoo Yoo, and Youngjung Uh. Exploiting spatial dimensions of latent in gan for real-time image editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[Li *et al.*, 2023] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot neural head avatar. *NeurIPS*, 2023.

[Lin *et al.*, 2024] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior, 2024.

[Lombardi *et al.*, 2018] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Trans. Graph.*, 37(4), jul 2018.

[Ma *et al.*, 2024] Haoyu Ma, Tong Zhang, Shanlin Sun, Xiangyi Yan, Kun Han, and Xiaohui Xie. Cvthead: One-shot controllable head avatar with vertex-feature transformer. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[Matsuda *et al.*, 2021] Nathan Matsuda, Brian Wheelwright, Joel Hegland, and Douglas Lanman. Reverse pass-through vr. In *ACM SIGGRAPH 2021 Emerging Technologies*, SIGGRAPH '21, New York, NY, USA, 2021. Association for Computing Machinery.

[Patel, 2024] Nilay Patel. Apple vision pro review: magic, until it's not, Januray 30 2024.

[Rambach *et al.*, 2020] Jason Rambach, Gergana Lilligreen, Alexander Schäfer, Ramya Bankanal, Alexander Wiebel, and Didier Stricker. A survey on applications of augmented, mixed and virtual reality for nature and environment, 2020.

[Richardson *et al.*, 2021] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[Schwartz *et al.*, 2020] Gabriel Schwartz, Shih-En Wei, Te-Li Wang, Stephen Lombardi, Tomas Simon, Jason Saragih, and Yaser Sheikh. The eyes have it: an integrated eye and face model for photorealistic facial animation. *ACM Trans. Graph.*, 39(4), aug 2020.

[Wang *et al.*, 2004] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2021] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[Wang *et al.*, 2024] Jingkai Wang, Jue Gong, Lin Zhang, Zheng Chen, Xing Liu, Hong Gu, Yutong Liu, Yulun Zhang, and Xiaokang Yang. Osdface: One-step diffusion model for face restoration, 2024.

[Wei *et al.*, 2019] Shih-En Wei, Jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. Vr facial animation via multiview image translation. *ACM Trans. Graph.*, 38(4), jul 2019.

[Wu *et al.*, 2024] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution, 2024.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers, 2021.

[Xie *et al.*, 2022] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

[Yue and Loy, 2024] Zongsheng Yue and Chen Change Loy. Difface: Blind face restoration with diffused error contraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9991–10004, 2024.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[Zheng *et al.*, 2023] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

[Zhu *et al.*, 2022] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022.